Virus Evolution, 2021, 7(1): veab015

doi: 10.1093/ve/veab015
Research Article

Evolution of DNA packaging in gene transfer agents

Emma S. Esterman,[1,†] Yuri I. Wolf,[2] Roman Kogay,[1] Eugene V. Koonin,[2,‡] and Olga Zhaxybayeva[1,3,*,§]

[1]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA, [2]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA and [3]Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA

*Corresponding author: E-mail: olga.zhaxybayeva@dartmouth.edu

[†]https://orcid.org/0000-0001-7365-0547

[‡]https://orcid.org/0000-0003-3943-8299

[§]https://orcid.org/0000-0002-1809-3909

Abstract

Gene transfer agents (GTAs) are virus-like particles encoded and produced by many bacteria and archaea. Unlike viruses, GTAs package fragments of the host genome instead of the genes that encode the components of the GTA itself. As a result of this non-specific DNA packaging, GTAs can transfer genes within bacterial and archaeal communities. GTAs clearly evolved from viruses and are thought to have been maintained in prokaryotic genomes due to the advantages associated with their DNA transfer capacity. The most-studied GTA is produced by the alphaproteobacterium *Rhodobacter capsulatus* (RcGTA), which packages random portions of the host genome at a lower DNA density than usually observed in tailed bacterial viruses. How the DNA packaging properties of RcGTA evolved from those of the ancestral virus remains unknown. To address this question, we reconstructed the evolutionary history of the large subunit of the terminase (TerL), a highly conserved enzyme used by viruses and GTAs to package DNA. We found that RcGTA-like TerLs grouped within viruses that employ the headful packaging strategy. Because distinct mechanisms of viral DNA packaging correspond to differences in the TerL amino acid sequence, our finding suggests that RcGTA evolved from a headful packaging virus. Headful packaging is the least sequence-specific mode of DNA packaging, which would facilitate the switch from packaging of the viral genome to packaging random pieces of the host genome during GTA evolution.

Key words: large terminase; TerL; DNA packaging; *Rhodobacter capsulatus*; RcGTA; alphaproteobacteria.

1. Introduction

Gene transfer agents (GTAs) are virus-like particles encoded and produced by certain bacteria and archaea (reviewed most recently by Lang, Westbye, and Beatty (2017) and Grull, Mulligan, and Lang (2018)). Unlike viruses, GTAs package fragments of the host genome instead of the genes that encode the GTA itself (Hynes et al. 2012). When GTA particles infect another cell, they can transfer the encapsidated genetic material to the recipient (Solioz, Yen, and Marris 1975; McDaniel et al. 2010; Hynes et al. 2012). The genomic loci that encode GTAs resemble prophages, indicating that GTAs evolved from viral ancestors. Although the function of GTAs is not firmly established, the prevailing hypothesis is that GTAs are not defective prophages, but instead are agents of horizontal gene transfer that are maintained in prokaryotic genomes due to the advantages associated with gene exchange, particularly, in stressful conditions (Lang, Westbye, and Beatty 2017; Kogay et al. 2020).

The best-studied GTA is produced by the alphaproteobacterium *Rhodobacter capsulatus*, and will be referred to as RcGTA. Production of the RcGTA particles is triggered by environmental

factors (Westbye et al. 2017a), occurs in a small fraction of the population (Fogg, Westbye, and Beatty 2012; Hynes et al. 2012), is regulated by host proteins (Westbye, Beatty, and Lang 2017b; Ding et al. 2019; Fogg 2019) and involves expression of genes that are found in at least five loci in the *R. capsulatus* genome (Hynes et al. 2016; Lang, Westbye, and Beatty 2017). The largest of these loci is a 17-gene 'head–tail cluster' that encodes proteins involved in head–tail morphogenesis and DNA packaging (Lang, Westbye, and Beatty 2017). There are homologs of the RcGTA head–tail cluster genes in other alphaproteobacteria (Shakya, Soucy, and Zhaxybayeva 2017), including several *Rhodobacterales* for which GTA production has been observed (Fu et al. 2010; Nagao et al. 2015). Additionally, homologs of head–tail cluster genes are present in numerous viruses and proviruses (Shakya, Soucy, and Zhaxybayeva 2017).

The small size of RcGTA and the low density of its packaged DNA precludes the particle from accommodating all of the genes required for its production (Bárdy et al. 2020). Instead, RcGTA packages seemingly random portions of the host genome (Hynes et al. 2012). In double-stranded DNA viruses of the realm *Duplodnaviria*, genome packaging is mediated by the terminase and portal proteins (Fokine and Rossmann 2014; Rao and Feiss 2015). Viral terminases typically consist of large and small subunits (Rao and Feiss 2008). The small subunit (TerS) binds to the DNA to be packaged and then recruits the large subunit (Casjens 2011; Rao and Feiss 2015). The large subunit (TerL), which consists of ATPase and nuclease domains, cuts the concatemeric viral DNA, translocates the DNA into the viral capsid with concomitant ATPase hydrolysis and, finally, cuts the DNA again to terminate packaging (Casjens 2011; Rao and Feiss 2015). Viruses evolved different strategies for packaging DNA into their capsids, and these strategies involve different classes of TerL (Casjens and Gilcrease 2009). Some viruses employ a headful packaging strategy whereby TerL initially cuts the viral concatemeric DNA at a specific sequence (*pac* site) and terminates packaging when the capsid is full, rather than at a particular sequence (Rao and Feiss 2008; Casjens and Gilcrease 2009). The virions produced by headful phages usually package more than 100 per cent of the phage genome length, and as a result, have terminally redundant, circularly permuted chromosomes (Casjens and Gilcrease 2009).

Viral TerLs with the same packaging mechanism tend to form clades in phylogenetic trees (Casjens et al. 2005). However, due to the diversity of TerL sequences, the relationships among the different functional classes of TerLs are not well-resolved (Casjens et al. 2005; Casjens and Gilcrease 2009; Merrill et al. 2016). Given its lack of sequence specificity, RcGTA is presumed to package fragments of the host DNA via the headful strategy (Casjens et al. 2005; Hynes et al. 2012). The RcGTA TerL and its alphaproteobacterial homologs formed a distinct group in previous phylogenies, but they did not cluster with or within the viral TerLs that are involved in headful packaging (Casjens et al. 2005; Casjens and Gilcrease 2009). Therefore, these phylogenies did not provide evidence of a headful packaging strategy in alphaproteobacterial GTAs, in part, due to limited sequence data available at the time.

In this study, we conducted a comprehensive evolutionary analysis of TerL sequences to better resolve the phylogenetic relationship of alphaproteobacterial RcGTA-like TerLs and viral TerLs with known packaging strategies. Our analyses suggest that RcGTA, and, by inference, the rest of the putative alphaproteobacterial GTAs, evolved from a virus that employed the headful packaging strategy. We also identified two amino acid substitutions that are conserved in the TerLs of the putative alphaproteobacterial GTAs and might be important for the DNA packaging properties of GTAs.

2. Methods

2.1 Retrieval and sequence-based clustering of large terminase homologs

Eighteen profiles covering one or both (ATPase and nuclease) domains of TerL were retrieved from the NCBI Conserved Domains Database (CDD) (Lu et al. 2020) (accessed on 11 December 2018). Two TerL profiles for distinct families of bacterial and archaeal viruses were added from Philosof et al. (2017) and Yutin et al. (2018). These 20 profiles (Supplementary Table S1) were used as queries for PSI-BLAST searches (E-value threshold of 0.01, effective database size of 2×10^7 sequences) (Altschul et al. 1997) against the NCBI non-redundant protein database (accessed in December 2018). Only the subject sequences that were taxonomically assigned to archaea, bacteria, and viruses were retained.

Partial TerL sequences were removed by ensuring the presence of both an ATPase (N-terminal) and a nuclease (C-terminal) domain using the following criteria: sequences either had to align to ≥75 per cent of a 'full' TerL profile that includes both TerL domains or align to different TerL profiles over their N- and C-terminal domains. A sequence was considered to align over a specific domain if it met one of two conditions: 1, if the sequence matched ≥75 per cent of an N- or C-terminal domain-specific profile and had at least 35 per cent of the protein length outside of the matched domain to contain the unmatched domain or 2, if a sequence aligned to just the N- or C-terminal portion of a 'full' TerL profile and had at least 35 per cent of the protein length outside of the matched domain to contain the unmatched domain.

The resulting 254,382 sequences were clustered using MMseqs2 (Steinegger and Söding 2017) with a similarity threshold of 0.75. From each of the obtained 11,298 clusters, a representative sequence of median length was selected for subsequent analyses.

2.2 Alignment of representative homologs and filtering out partial sequences

The representative TerL sequences were iteratively aligned and clustered using the approach described by Wolf et al. (2018). Briefly, the sequences were clustered with a similarity threshold of 0.5 using UCLUST (Edgar 2010). The clustered sequences were aligned using MUSCLE (Edgar 2004) and alignment sites that contained more than 67 per cent gaps were temporarily removed. Pairwise similarity scores between cluster alignments were calculated using HHSEARCH (Söding 2005), converted to a distance matrix and used to build a UPGMA tree (Sokal and Michener 1958). All of the branches of the UPGMA tree above a depth threshold of 2.3 were used to guide progressive alignment of the clusters using HHALIGN (Söding 2005). The removed sites were reinserted back into their original sequences after the profile–profile alignment. These alignment and clustering steps were repeated for 20 iterations, when 11,230 of the sequences formed one alignment.

Of the remaining 68 sequences that failed to align, two were clearly TerLs but contained inteins, which were manually removed. Five other sequences were also likely TerLs because they were longer than 300 amino acids, exhibited significant similarity to a TerL profile via CDD searches (E-value <0.001),

and contained recognizable Walker A motif and nuclease catalytic residues. The seven sequences were profile-aligned to the alignment of 11,230 sequences using more relaxed criteria (similarity threshold of 0.01 and UPGMA depth threshold of 6). The remaining 61 sequences did not meet these criteria and were discarded.

The new alignment of 11,237 sequences contained partial sequences that lacked a Walker A motif. To remove these, each sequence's similarity was scored to the alignment's consensus sequence using a BLOSUM62 substitution matrix and the score was compared to the score of sequences with 100 per cent identity to the consensus sequence. Sequences with a score less than 10 per cent of the perfect match score were removed. Then, the alignment was used as a PSSM in a PSI-BLAST search against all of the sequences within the alignment. Only the sequences that matched to ≥75 per cent of the PSSM were retained. The sections of the 11,060 sequences that passed this criterion were extracted and re-aligned using the above-described iterative alignment procedure with a clustering similarity threshold of 0.5 and UPGMA depth threshold of 2.3. After 23 iterations of alignment and clustering, 11,057 of the sequences aligned. The three sequences that did not align were longer than 300 amino acids, exhibited significant similarity to a TerL profile via CDD searches (E-value <0.001), and contained a recognizable Walker A motif and nuclease catalytic residues. Therefore, they were retained and aligned to the main alignment using more relaxed parameters of a clustering similarity threshold of 0.01 and UPGMA depth threshold of 6.

2.3 Alignment trimming

The alignment of 11,060 sequences was trimmed to remove all columns with more than 50 per cent gaps and less than 10 per cent amino acid homogeneity. The homogeneity value of an alignment column was defined and calculated using the following procedure. For each of the N=11,060 sequences, column-based sequence weights w_i $\left(\sum_{i=1}^{N} w_i = 1\right)$ were assigned according to Henikoff and Henikoff (1994). The score of an alignment column against an amino acid x was calculated as $Q_x = \sum_{i=1}^{N} w_i S_{a_i,x}$, where a_i is an amino acid in the i-th sequence and $S_{a_i,x}$ is the BLOSUM62 substitution matrix score for a pair of amino acids a_i and x (Henikoff and Henikoff 1993). As the consensus amino acid of the column c, the amino acid with the highest score Q_c, that is, $c = \mathrm{argmax}_x\, Q_x$, was selected. An expectation of the score of the given alignment column against a randomly selected amino acid R was calculated as $Q_R = \sum_b f_b Q_b$, where f_b is the vector of relative frequencies of amino acids ($\sum_b f_b = 1$, $b \in \{\mathrm{Ala}, \mathrm{Tyr}\}$). The homogeneity of an alignment column was defined as $H = \max\left(\frac{Q_c - Q_R}{S_{c,c} - Q_R}, 0\right)$. The homogeneity ranges from 0 (the alignment column score does not exceed the random expectation score Q_R) to 1 (the alignment column score is equal to the maximum possible score $S_{c,c}$).

2.4 Reconstruction of large terminase phylogeny

The trimmed alignment of 11,060 sequences was used to reconstruct an initial phylogenetic tree in FastTree v. 2.1.4 (Price, Dehal, and Arkin 2010) using the Whelan and Goldman (WAG) substitution model (Whelan and Goldman 2001) and 20 gamma-distributed rate categories (Yang 1994). The initial tree was used as a guide tree to refine our alignment, which in turn was used to reconstruct an improved tree. To this end, the sequences were divided into two sets, those that formed distinct groups on the tree and the remaining ones. The sequences that formed distinct groups were re-aligned using a clustering similarity threshold of 0.01 and UPGMA depth threshold of 2.3, whereas the other sequences were aligned using more stringent parameters (similarity threshold of 0.66 and UPGMA depth threshold of 1.3). These alignments were profile-aligned using a similarity threshold of 0.5 and UPGMA depth threshold of 2.3. Nine sequences did not join the main alignment and were discarded due to low scores against the consensus, calculated with a BLOSUM62 matrix as described above. The resulting alignment of 11,051 sequences was trimmed to remove all columns with more than 50 per cent gaps and less than 10 per cent amino acid homogeneity, as defined above. The final phylogenetic tree of 11,051 TerL homologs (Fig. 1) was reconstructed using FastTree v. 2.1.4 (Price, Dehal, and Arkin 2010) with the WAG substitution model (Whelan and Goldman 2001) and 20 gamma-distributed rate categories (Yang 1994).

2.5 Identification of RcGTA-like large terminases

In the phylogenetic tree of TerLs (Fig. 1), a group of 616 TerLs was labeled as the 'RcGTA-like containing group'. This group includes 507 TerLs of the 526 RcGTA-like TerLs that were identified and curated by Kogay et al. (2019). Nineteen RcGTA-like TerLs from the dataset of Kogay et al. (2019) are absent in our dataset because they were not in GenBank at the time of our data collection (December 2018). The 616 TerLs were classified as either 'RcGTA-like' or 'virus-like' using a machine learning approach implemented in the GTA-Hunter program (Kogay et al. 2019).

2.6 Examination of the genomic neighborhoods of large terminases

For the 604 TerLs within the 'RcGTA-like containing' group (Fig. 1) that originated from bacterial and archaeal genomes, the presence of 11 other RcGTA-like genes near the *terL* gene was examined. To this end, the RcGTA-like genes from the training set of Kogay et al. (2019) were used as queries in a BlastP search against the assemblies of the bacterial and archaeal genomes (E-value <0.001; query and subject overlap by at least 60 per cent of their length) (Altschul et al. 1997). The detected RcGTA gene homologs were classified as 'RcGTA-like' or 'virus-like' using GTA-Hunter (Kogay et al. 2019). The detected RcGTA gene homologs were also clustered into regions using DBSCAN, with a maximum distance cutoff of 8,000 bp between adjacent genes (Ester et al. 1996; Kogay et al. 2019). If a *terL* gene was embedded in a region containing at least 6 of the 11 RcGTA-like genes, it was classified as being in a 'large RcGTA-like element'. If a *terL* gene was located in a region containing 1 to 5 RcGTA-like genes, it was classified as being in a 'small RcGTA-like element'. The classification of the sequence represented in the phylogeny was assumed to be the same for the rest of the cluster members although this was not directly verified.

2.7 Assignment of packaging strategy to viral large terminases

A list of viruses with experimentally determined packaging mechanisms was compiled from the phylogenetic tree of Casjens and Gilcrease (2009). Viruses from the phylogenetic tree of Merrill et al. (2016) were also added, for most of which experimental evidence of the packaging mechanism is available. The dataset of

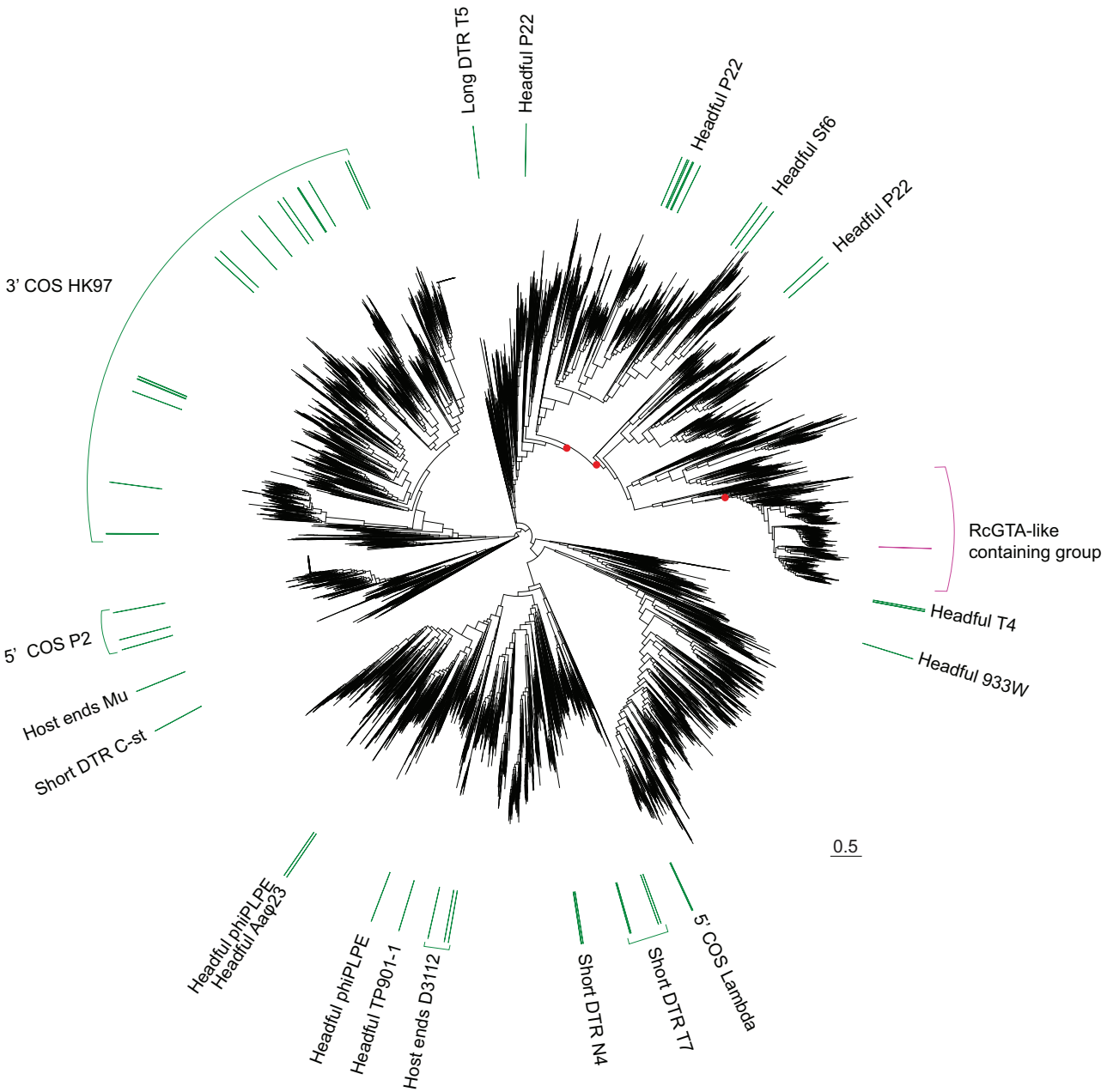

Figure 1. Phylogeny of TerLs from 11,051 viruses, prophages, and GTAs. The TerL protein from RcGTA is denoted by a pink bar. The pink bracket outlines a subtree that contains RcGTA-like TerLs and is shown in detail in Fig. 2. The green bars denote the viruses that have experimentally determined packaging strategies (Supplementary Table S2). The viruses with experimentally determined packaging strategies are labeled by the packaging strategy (Headful, Cohesive ends [COS], Direct Terminal Repeats [DTR], and Host Ends) followed by a prototype phage from that group (e.g., P22). Support values (aLRT) of >75 per cent denoted by red dots are only shown for a selection of branches relevant to grouping RcGTA-like TerLs within phages that employ a headful DNA packaging strategy. Scale bar, amino acid substitutions per site. Tree topology with all support values is available in NEWICK format in Supplementary Data. The patterns of this phylogeny are consistent with those in the phylogenies reconstructed using a more accurate maximum likelihood inference carried out using the IQ-TREE program (see Supplementary Dataset); in particular, the RcGTA-like TerLs continue to form a well-supported group (100 per cent of bootstrap samples), which forms a sister group to TerLs of phages that utilize a headful packaging strategy (75 per cent of bootstrap samples).

the 252,614 TerL homologs was searched for these 87 viruses using TerL accession numbers provided by Merrill et al. (2016) and NCBI taxonomy IDs for the viruses from Casjens and Gilcrease (2009). Of the 87 viruses, 73 were present in our dataset (Supplementary Table S2). Due to the close sequence similarity, some of the 73 TerLs belong to the same MMSEQ clusters, and therefore are represented by 58 TerLs on our phylogeny of 11,051 TerLs (Fig. 1). The 58 representative viruses for which the packaging mechanism was not known were assigned the mechanism of a virus from the same cluster with a known packaging strategy, under the assumption that the similarity of their TerL amino acid sequences is sufficient to imply the same packaging mechanism.

2.8 Validation of the reconstructed phylogenetic patterns with more accurate maximum likelihood analyses

To confirm that the phylogenetic relationships obtained from the FastTree program (Price, Dehal, and Arkin 2010) were not

impacted by its limited tree search and optimization capabilities, additional phylogenetic trees were reconstructed using IQ-TREE v 1.6.7 (Nguyen et al. 2015) from two datasets subsampled from the 11,051 TerLs. The first dataset of 342 TerLs was constructed to broadly represent the TerL diversity (Fig. 1). The dataset contains the 58 representative viral TerLs (described in Section 2.7), 50 TerLs randomly sampled from group 1 of the 'RcGTA-like containing group' (Fig. 2), all TerLs from group 2, 70 TerLs from group 3, 50 TerLs randomly sampled from the rest of the 'RcGTA-like containing group', and 100 randomly sampled TerLs from the rest of the whole TerL tree (Fig. 1). The second dataset of 346 TerLs was constructed to represent well the TerLs from the 'RcGTA-like containing group' (Figs 1 and 2). The dataset contains 12 representative viral TerLs with either P22 or Sf6-like headful packaging strategies, 70 TerLs randomly sampled from group 1, all TerLs from group 2, and 250 TerLs randomly sampled from group 3. For both datasets, the aligned sequences were retrieved from the trimmed alignment of 11,051 TerLs (see Section 2.4) and gap-only sites were removed. The optimal evolutionary model was selected using ModelFinder (Kalyaanamoorthy et al. 2017), as implemented in IQ-TREE. Support values for branches were calculated using ultrafast bootstrap approximation with 1,000 replicates (Hoang et al. 2018), as implemented in IQ-TREE. The tree reconstructed from the second dataset was rooted with the headful P22 viral TerL that, out of the headful P22 viral TerLs in Fig. 1, is the most distantly related to the 'RcGTA-like containing group'.

2.9 Annotation of prophages in regions encoding virus-like TerLs

To examine whether the bacterial TerLs classified as virus-like reside in prophages, prophages were predicted in the corresponding nucleotide genome sequences using PHASTER (Arndt et al. 2016, accessed in May 2020). The regions labeled as intact (score >90) or questionable (score 70–90) prophages were retained, while the regions labeled as incomplete prophages (score <70) were discarded. One region surrounding the virus-like TerL (accession WP_020474221) was predicted by PHASTER to be an intact prophage and was also classified as a small RcGTA-like element by GTA-Hunter, due to the presence of one RcGTA-like gene. The PHASTER prediction of this region as a prophage was considered to supersede the small RcGTA-like element classification.

2.10 Detection of conserved sites that differentiate RcGTA-like and virus-like TerLs

The amino acid sequences of the 616 TerLs in the subtree shown on Fig. 2 were re-aligned using MAFFT-linsi v. 7.305 (Katoh and Standley 2013). The alignment was scanned for sites conserved in more than 80 per cent of the TerL homologs from group 1 (virus-like) or group 3 (RcGTA-like) (Fig. 2). Of these detected sites, only the sites with at least 70 per cent between-group difference in the relative abundance of the most conserved amino acid were retained.

The positions of the two identified sites relative to the known TerL structural domains were determined by searching CDD (Lu et al. 2020) with RcGTA TerL RefSeq record WP_031321187 as a query (database accessed on 26 July 2020). The conservation of the two detected sites within the 'RcGTA-like containing group' was visualized using a subset of the 616 TerLs (Fig. 2): 15 randomly sampled TerLs from group 1, all

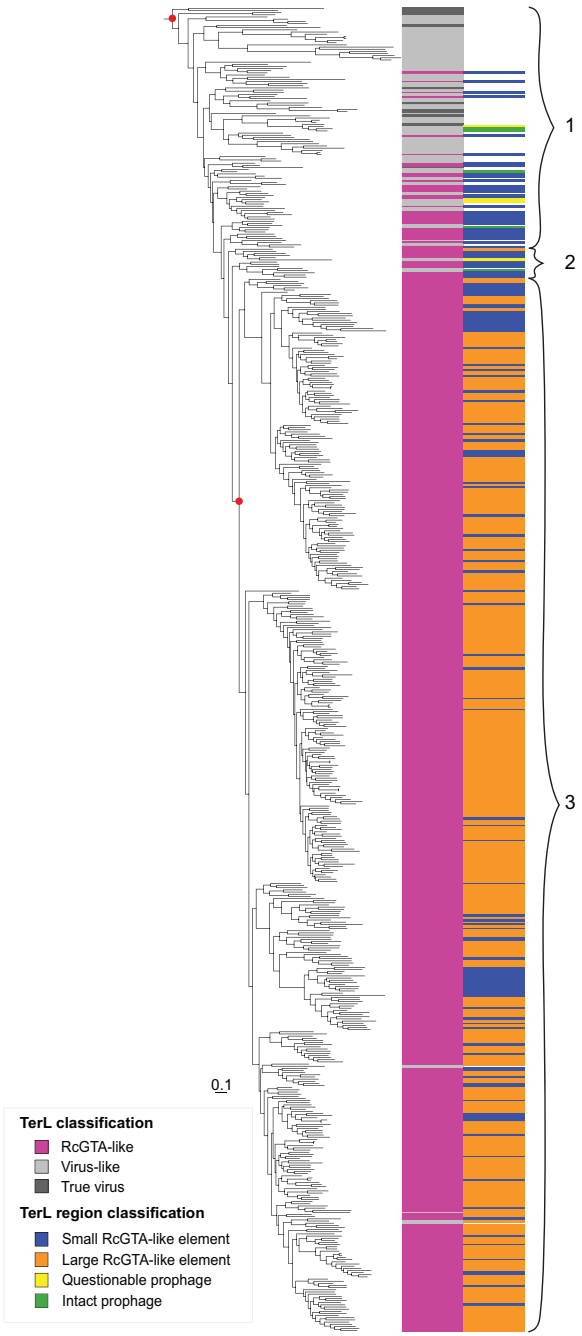

Figure 2. A subtree of the phylogeny shown on Fig. 1 that contains RcGTA-like TerLs. Bars in the first column next to the branches of the subtree indicate whether the TerLs are from true viruses or, if they are found in bacterial or archaeal genomes, whether they were classified by GTA-Hunter as 'RcGTA-like' or 'Virus-like'. Bars in the second column denote whether the genomic neighborhood of the *terL* gene contains at least six RcGTA-like genes ('large RcGTA-like element'), between one and five RcGTA-like genes ('small RcGTA-like element'), a questionable prophage or an intact prophage. TerLs without a colored bar in the second column were not predicted as being in a prophage or RcGTA-like element. Group 1 includes mostly virus-like TerLs as well as TerLs from predicted prophages and true viruses. Group 2 contains a mixture of TerLs from a large element, small RcGTA-like elements, and predicted prophages. Group 3 mostly contains RcGTA-like TerLs from 'true' GTAs (large RcGTA-like elements). Red dots indicate two nodes that are relevant for the grouping of the entire subtree and for the TerLs in group 3, and have aLRT support values of >75 per cent. The patterns of this phylogeny are consistent with those in the phylogeny reconstructed using IQ-TREE (see Supplementary Data); in particular, the node corresponding to group 3 has 96 per cent bootstrap support. Scale bar, amino acid substitutions per site.

TerLs from group 2, 14 randomly sampled TerLs from group 3, and a representative TerL from the cluster that contains the RcGTA TerL. The aligned TerLs were retrieved from the alignment of 616 TerLs, and the gap-only alignment positions were removed. Secondary structure information was obtained via HHPred using RcGTA TerL as a query (Zimmermann et al. 2018).

The locations of the two sites were also visualized on a 3D structure of TerL from the *Shigella* phage Sf6 (PDB ID 4IDH) (Zhao et al. 2013) which, based on our phylogenetic inference, is the TerL most closely related to RcGTA-like TerLs for which a structure is available. The homologous positions of the substitutions in the *Shigella* phage Sf6 TerL were identified by aligning it to the RcGTA TerL using HHPred (Zimmermann et al. 2018). The visualization was carried out in PyMOL v 2.4 (Schrödinger, LLC 2020).

3. Results

3.1 RcGTA-like TerLs belong within the group of TerLs of headful packaging phages

Of the 11,051 representative TerL homologs from bacteria, archaea, and viruses, 616 are closely related to the RcGTA TerL and form a well-supported group in the phylogenetic tree (Fig. 1). Of these 616 TerLs, twelve are encoded in viral genomes, whereas the remaining 604 are found in 601 bacterial and 3 archaeal genomes. Using a machine learning approach that relies on amino acid composition, we classified the 604 bacterial and archaeal TerL homologs as either 'RcGTA-like' (527) or 'virus-like' (77) (Supplementary Table S3). By mapping 73 TerLs with experimentally determined packaging strategies onto the phylogeny, we found that RcGTA-like TerLs fall, with strong support, within a group of headful packaging phages (Fig. 1). Therefore, our phylogeny implies that the RcGTA-like TerLs evolved from a viral TerL that employed a headful packaging strategy and is thus consistent with the earlier proposed hypothesis that RcGTA-like TerLs use a headful mechanism to package host DNA (Casjens et al. 2005; Hynes et al. 2012). Of the TerLs from the viruses with experimentally determined packaging strategies, *Enterobacteria* phage P22-like TerLs are the closest relatives of the RcGTA-like TerLs. This affinity contrasts the previous results, from analyses of much smaller data sets, according to which T4-like (Lang and Beatty 2000; Hynes et al. 2012) or T7-like (Casjens et al. 2005) TerLs were found to be most closely related to the RcGTA-like TerLs.

3.2 Phylogenetic evidence of a single origin of GTA TerLs

Whereas the TerLs of viruses that employ headful DNA packaging specifically package the viral genome into the capsid, RcGTA TerL lacks sequence specificity and packages random segments of the bacterial genome (Hynes et al. 2012). To evaluate if non-specific DNA packaging evolved once or multiple times, we first sought to determine more accurately which of the RcGTA-like TerLs likely belong to *bona fide* GTAs. Shakya, Soucy, and Zhaxybayeva (2017) hypothesized that genomic regions with a smaller number of recognizable RcGTA gene homologs are more likely to be prophages than GTAs because these regions tend to have a more virus-like GC content relative to their host, evolve faster and are more often associated with viral genes. Therefore, some of the TerLs classified as 'RcGTA-like' might not belong to RcGTA-like elements in cases when the alphaproteobacterial genomes that contain these genes lack homologs of other RcGTA genes. Among the 527 RcGTA-like TerLs, we classified 391 as 'large' (containing at least six RcGTA-like genes near the *terL* gene) and 136 as 'small' (1–5 RcGTA-like genes) elements (Fig. 2 and Supplementary Table S3).

Within the subtree that contains RcGTA-like TerLs (Fig. 2), all but one of the TerLs found in large elements form a well-supported clade (group 3 in Fig. 2). The one TerL from a large element that falls outside this clade is a representative of a cluster of three TerLs that are found in the genomes of alphaproteobacteria *Zavarzinia compransoris* DSM 1231, *Zavarzinia* sp. HR-AS and *Oleomonas* sp. K1W22B-8. This TerL belongs to a narrow 'transition zone' (group 2 in Fig. 2) between the group 3 TerLs and the deepest branches of the subtree that include exclusively viral and 'virus-like' sequences (group 1 in Fig. 2). The transition zone also contains a mix of RcGTA-like TerLs from small elements and virus-like TerLs, including the TerL from the intact prophage predicted in the genome of a planctomycete *Zavarzinella formosa* DSM 19928. None of the TerLs within this transition zone come from functionally characterized viruses or GTAs. Thus, the phylogeny indicates that RcGTA-like TerLs likely evolved only once from a viral TerL, in an ancestor of group 3, by acquiring the capability to package DNA non-specifically. The positions of the TerLs from *Zavarzinia*'s and *Oleomonas*' putative GTAs and the *Zavarzinella* prophage could be explained by horizontal gene transfer, as previously documented in some instances for other RcGTA-like genes (Yang et al. 2017) and discussed in detail below.

3.3 Viruses might mediate horizontal gene exchange of RcGTA-like genes

In addition to the above-discussed predicted prophage from *Zavarzinella formosa* DSM 19928, we identified two intact prophages in the genomes of the firmicute *Thermoactinomyces* sp. DSM 45892 and the alphaproteobacterium *Methylobacterium terrae* 17Sr1-28, and 16 viruses that encode TerLs that are phylogenetically most closely related to the RcGTA-like TerLs (**Tables 1 and 2**). Notably, the TerLs of these three prophages are even more closely related phylogenetically to the TerLs from large elements than the 16 viruses are (Fig. 2), but whether they produce functional virions is unknown.

Some of the 16 viruses encoding TerLs related to GTA TerLs infect GTA-containing alphaproteobacteria and possess genes that are more closely related to GTA genes than to their homologs in other viruses. For example, *Dinoroseobacter* phage vB_DshS-R5C contains homologs of four putative tail genes of the RcGTA (genes *g12–g15*; Yang et al. 2017). The predicted prophage in *Zavarzinella formosa* DSM 19928 encompasses a homolog of the adaptor gene (*g6*) that is RcGTA-like in amino acid composition. These observations suggest an ongoing exchange and recombination of RcGTA-like genes between viruses and alphaproteobacteria, which likely explains the presence of virus-like TerLs within groups 2 and 3 (Fig. 2). The TerL phylogeny also indicates that such gene exchange might extend beyond alphaproteobacteria because at least four bacterial TerLs within groups 2 and 3 come from non-alphaproteobacterial genomes (OYV96073, WP_020474221, WP_110156686, and OQX66442). Because the viruses with known hosts infect a wide range of bacteria that live in environments similar to those occupied by GTA-containing alphaproteobacteria (**Table 1**), they either might have an opportunity for gene exchange with viruses that infect GTA-containing bacteria or might be capable of infecting GTA-containing bacteria, in addition to their currently known hosts.

Table 1. Sixteen viruses with TerLs most closely related to RcGTA-like TerLs.

Virus	TerL accession number	On tree?[a]	Host name	Host taxonomic class	Habitat	GTA?[b]	Reference[c]
Arthrobacter phage Tank	ALY10550.1	Yes	*Arthrobacter* sp. ATCC 21022	Actinobacteria	Soil	No	GenBank[d]
Arthrobacter phage Wilde	ALY10802.1	No	*Arthrobacter* sp. ATCC 21022	Actinobacteria	Soil	No	GenBank
Caulobacter phage Sansa	AKU43425.1	Yes	*Caulobacter crescentus* CB15	Alphaproteobacteria	Aquatic	Yes	26450723
Colwellia phage 9A	AFK66668.1	Yes	*Colwellia psychrerythraea* 34H	Gammaproteobacteria	Cold environments	No	23224375
Dinoroseobacter phage vB_DshS-R5C	ARB06077.1	Yes	*Dinoroseobacter shibae* DFL12T	Alphaproteobacteria	Ocean surface	Yes	28505134
Gordonia phage GMA2	AKJ72540.1	Yes	*Gordonia malaquae* A448	Actinobacteria	Activated sludge	No	26241321
Microbacterium phage Hyperion	AWN03535.1	Yes	*Microbacterium foliorum* NRRL B-24224	Actinobacteria	Soil	No	GenBank
Microbacterium phage OneinaGillian	AYB70129.1	Yes	*Microbacterium foliorum* NRRL B-24224	Actinobacteria	Soil	No	GenBank
Microbacterium phage Squash	AWN04641.1	No	*Microbacterium foliorum* NRRL B-24224	Actinobacteria	Soil	No	GenBank
Nitrincola phage 1M3-16	AHX01069.1	Yes	*Nitrincola sp.* 1M3-16	Gammaproteobacteria	Hypersaline alkaline lake	No	GenBank
Salinibacter virus M1EM-1	AUO78912.1	No	*Salinibacter ruber* M1	Bacteroidetes	Saltern	No	29099492
Salinibacter virus M8CR30-2	AUO79033.1	Yes	*Salinibacter ruber* M8	Bacteroidetes	Saltern	No	29099492
Salinibacter virus M8CR30-4	AUO79074.1	No	*Salinibacter ruber* M8	Bacteroidetes	Saltern	No	29099492
Streptomyces phage mu1/6	ABD94197.1	Yes	*Streptomyces aureofaciens*	Actinobacteria	Soil	No	18062183
Environmental Halophage eHP-25	AFH22435.1	Yes	Unknown, hypothesized to be Nanohaloarchaea	Unknown	Saltern	No	22479446
Uncultured Mediterranean phage uvDeep-CGR2-KM21-C338	ANS03529.1	Yes	Unknown	Unknown	Deep ocean	No	27460793

[a]Whether the TerL is the one present on the tree or clustered with one of the other viral TerLs due to high sequence similarity.
[b]Whether the host's genome contains an RcGTA-like element.
[c]PubMed ID of the paper that discusses the isolation and/or genome of the virus.
[d]Direct submission to GenBank.

3.4 Two amino acid changes distinguish GTA and viral TerLs

Although viral TerLs that are most closely related to RcGTA-like TerLs have not been experimentally characterized, examination of amino acids that are conserved in RcGTA-like TerLs from large elements (group 3 on Fig. 2) but not in closely related viral TerLs (group 1 on Fig. 2), or vice versa, might help pinpoint the changes that are important for the unique packaging strategy of GTAs. We did not identify any amino acids that are conserved in group 1 TerLs but not in group 3 TerLs, but found two amino acids (located at positions 282 and 292 in the RcGTA TerL; RefSeq record WP_031321187) that are conserved in the group 3 TerLs but not in the group 1 TerLs (Supplementary Table S4 and Fig. S1). In position 292, 99 per cent of the group 3 TerLs, but only 4 per cent of the group 1 TerLs, contain cysteine, whereas 59 per cent of the group 1 TerLs contain threonine. However, given that the threonine to cysteine substitution results in a reduction of the number of carbons per side chain, selection for the reduction in the energetic cost of GTA production (Kogay et al. 2020) cannot be excluded as a driver for this substitution. In position 282, 90 per cent of the group 3 TerLs but no group 1 TerLs contain proline, whereas 64 per cent of the group 1 TerLs but only 6 per cent of the group 3 TerLs contain alanine. Proline contains two more carbons in its side chain than alanine, and therefore, this substitution cannot be selected for energetic cost savings. In TerLs from *bona fide* GTAs of *R. capsulatus* and *Dinoroseobacter shibae*, cysteine is found at position 292 in both proteins, whereas in position 282 the *Dinoroseobacter shibae* TerL has proline and RcGTA TerL has alanine.

Within the TerL protein structure, the two amino acids are located in the nuclease domain (Rao and Feiss 2015) and lie in close proximity at the opposite ends of a loop that extends toward the translocating DNA (Supplementary Fig. S2 and Figure 3B in Zhao et al. (2013)). The importance of these residues with respect to the functionality of the GTA TerLs remains to be elucidated.

Table 2. Three predicted intact prophages with TerLs closely related to RcGTA-like TerLs.

Host name	TerL accession number	Predicted prophage coordinates in the host's genome	Host taxonomic class	Host habitat	Reference[a]
Zavarzinella formosa DSM 19928	WP_020474221.1	NZ_JH636446.1: 39426–61037	Planctomycetia	Wetlands	22740668
Methylobacterium terrae 17Sr1-28	WP_109959484.1	NZ_CP029553.1: 2863294–2900746	Alphaproteobacteria	Soil	31463788
Thermoactinomyces sp. DSM 45892	SDY22851.1	FNPL01000003.1:4278–49984	Bacilli	Not Provided	GenBank[b]

[a]PubMed ID of the paper that discusses the isolation and/or genome of the host organism.
[b]Direct submission to GenBank.

4. Discussion

RcGTA is hypothesized to package DNA via a headful mechanism because RcGTA particles encapsidate random pieces of *R. capsulatus*' DNA, which would likely be facilitated by a non-sequence-specific TerL (Casjens et al. 2005; Hynes et al. 2012). Previous experiments support the hypothesis that RcGTA utilizes headful packaging because the packaged DNA fragments have different sequences at the ends (Hynes et al. 2012). The large dataset of available TerL sequences allowed us to obtain phylogenetic evidence that the RcGTA-like TerLs evolved from the large terminases of headful-packaging phages (**Fig. 1**). Our findings suggest that RcGTA either continues to employ the headful packaging strategy of its ancestor or modified it into a unique DNA packaging strategy.

Previous studies have reported that the RcGTA TerL was most closely related either to the TerLs of T7-like viruses, which use a sequence-specific packaging mechanism (Casjens et al. 2005), or to the TerLs of T4-like viruses, which employ the headful packaging mechanism (Lang and Beatty 2000; Hynes et al. 2012). However, we found that the RcGTA-like TerLs are more similar to the TerLs of headful-packaging P22-like viruses. This discrepancy is likely due to the vastly expanded set of viral sequences now available in GenBank and the more sensitive search method that we used to identify viral TerL homologs.

In further support of the origin of RcGTA from a virus that employed a headful packaging mechanism, several structural proteins of RcGTA have the highest sequence and secondary structure similarity to the corresponding proteins in viruses that also utilize a headful packaging strategy (Bárdy et al. 2020). Specifically, the tail tape measure protein of RcGTA is homologous to the tail-needle protein of bacteriophage P22, domains of the RcGTA hub and megatron proteins are homologous to their counterparts in bacteriophage T4, and the RcGTA stopper and tail terminator proteins are homologous to those from bacteriophage SPP1 (Bárdy et al. 2020).

We identified TerLs from several viruses and predicted prophages that are phylogenetically closer relatives of the RcGTA TerL than P22-like TerLs. These viruses and predicted prophages either infect alphaproteobacteria or at least are found in the same environments as GTA-containing alphaproteobacteria. Because the specific mechanisms of headful packaging differ among phages (Casjens et al. 1992, 2004; Bhattacharyya and Rao 1993), experimental characterization of packaging in viruses that are closely related to GTAs could offer further insight into the origin of the GTA TerLs and their packaging mechanism.

The TerL phylogeny presented here supports the single origin of the RcGTA head–tail cluster in alphaproteobacteria because large RcGTA-like elements grouped together. With the newfound support for a single origin of RcGTA-like TerLs from a TerL of a headful-packaging phage, we propose that a headful-packaging TerL in the RcGTA ancestor underwent key changes that resulted in the switch from packaging the GTA genome to packaging random, small pieces of the host genome (Hynes et al. 2012) with a substantially lower density of DNA in the capsid (Bárdy et al. 2020). The selection for reduced energy cost of GTA protein production that apparently occurred after the origin of RcGTA-like elements in alphaproteobacteria (Kogay et al. 2020) makes it challenging to pinpoint amino acid changes in GTA TerLs that contribute to this transition.

The loss of specificity for the GTA genome and the reduction in the DNA packaging density rule out self-propagation of the GTA genome mediated by virions. Strikingly, the RcGTA genes appear to be actively precluded from packaging, being the least frequently packaged region in the alphaproteobacterial genome that is incorporated into the GTA particles (Hynes et al. 2012). The mechanism behind this exclusion is unknown, one possibility being that intensive expression of the RcGTA genes interferes with their packaging (Hynes et al. 2012).

In addition to TerL, other GTA proteins that are involved in DNA packaging, including TerS and portal, might contribute to the unique DNA packaging features of RcGTA. In particular, TerS proteins determine where packaging initiates and control the specificity of packaging through the recognition of *pac* sites (Leavitt et al. 2013). However, no discrete packaging start sites were found in the RcGTA genome (Hynes et al. 2012). The RcGTA gene *g1*, which is adjacent to the *terL* gene *(g2)* in the RcGTA genome, has been recently shown to encode TerS (Sherlock, Leong, and Fogg 2019). Sherlock, Leong, and Fogg (2019) demonstrated that the RcGTA TerS binds non-specifically to DNA with low affinity due to the absence of a specific DNA-binding domain and the retention of non-specific DNA binding activity. A TerS protein with altered DNA-binding characteristics and a modified headful TerL might together underlie the random packaging that is characteristic of RcGTA.

Acknowledgements

We thank Zhengshuang Hua for insightful discussions and help with using Dartmouth computing facilities.

Funding

This work was supported by the following awards from Dartmouth College to E.S.E.: Sophomore Research Scholarship, James O. Freedman Presidential Scholarship, Thomas B. Roos Memorial Fund Fellowship, and a Kaminsky Undergraduate Research Award. Additionally, this work was

supported by an Intramural Research and Training Award from the National Institutes of Health to E.S.E., by the Simons Foundation Investigator in Mathematical Modeling of Living Systems award #327936 to O.Z., by the National Science Foundation award DEB-1551674 to O.Z., and by the Intramural Research Program of the U.S. National Institutes of Health (National Library of Medicine) to Y.I.W. and E.V.K.

Data Availability

The following data is provided in our Supplementary Data available via FigShare (https://doi.org/10.6084/m9.figshare.12191691): GenBank accession numbers of the 254,382 RcGTA TerL protein homologs that are taxonomically assigned to bacteria, archaea, or viruses and likely include both ATPase and nuclease domains, GenBank accession numbers of the 252,614 RcGTA TerL protein homologs that are represented by 11,051 TerLs in the tree in **Fig. 1**, GenBank accession numbers of the 11,051 amino acid sequences used for reconstruction of the tree in **Fig. 1**, alignment of 11,051 TerL amino acid sequences (untrimmed and trimmed), phylogenetic tree of 11,051 TerLs shown in **Fig. 1**, alignment of amino acid sequences of 616 TerLs from the subtree shown in **Fig. 2**, alignments and phylogenetic trees of 342 and 346 TerLs used in IQ-TREE phylogenetic reconstructions. All alignments are in FASTA format and all phylogenetic trees are in NEWICK format.

Supplementary data

Supplementary data are available at *Virus Evolution* online and in FigShare repository at https://doi.org/10.6084/m9.figshare.12191691.

Conflict of interest: None declared.

Author Contributions

E.S.E., O.Z., Y.I.W., and E.V.K. designed the study. E.S.E. collected data. E.S.E. and R.K. performed the analyses. E.S.E., O.Z., Y.I.W., R.K., and E.V.K. interpreted the results. E.S.E. and O.Z. wrote the initial draft of the manuscript. E.S.E., O.Z., Y.I.W., R.K., and E.V.K. revised the manuscript.

References

Altschul, S. F. et al. (1997) 'Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs', *Nucleic Acids Research*, 25: 3389–402.

Arndt, D. et al. (2016) 'PHASTER: A Better, Faster Version of the PHAST Phage Search Tool', *Nucleic Acids Research*, 44: W16–21.

Bárdy, P. et al. (2020) 'Structure and Mechanism of DNA Delivery of a Gene Transfer Agent', *Nature Communications*, 11: 3034.

Bhattacharyya, S. P., and Rao, V. B. (1993) 'A Novel Terminase Activity Associated with the DNA Packaging Protein gp17 of Bacteriophage T4', *Virology*, 196: 34–44.

Casjens, S. R. (2011) 'The DNA-Packaging Nanomotor of Tailed Bacteriophages', *Nature Reviews Microbiology*, 9: 647–57.

——, and Gilcrease, E. B. (2009) 'Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions', *Methods in Molecular Biology (Clifton, N.J.)*, 502: 91–111.

Casjens, S. et al. (1992) 'Molecular Genetic Analysis of Bacteriophage P22 Gene 3 Product, a Protein Involved in the Initiation of Headful DNA Packaging', *Journal of Molecular Biology*, 227: 1086–99.

—— et al. (2004) 'The Chromosome of *Shigella flexneri* Bacteriophage Sf6: Complete Nucleotide Sequence, Genetic Mosaicism, and DNA Packaging', *Journal of Molecular Biology*, 339: 379–94.

Casjens, S. R. et al. (2005) 'The Generalized Transducing *Salmonella* Bacteriophage ES18: Complete Genome Sequence and DNA Packaging Strategy', *Journal of Bacteriology*, 187: 1091–104.

Ding, H. et al. (2019) 'Induction of *Rhodobacter capsulatus* Gene Transfer Agent Gene Expression is a Bistable Stochastic Process Repressed by an Extracellular Calcium-Binding RTX Protein Homologue', *Journal of Bacteriology*, 201: e00430–19.

Edgar, R. C. (2004) 'MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity', *BMC Bioinformatics*, 5: 113.

—— (2010) 'Search and Clustering Orders of Magnitude Faster than BLAST', *Bioinformatics*, 26: 2460–1.

Ester, M. et al. (1996). 'A density-based algorithm for discovering clusters in large spatial databases with noise' in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–31. Portland: AAAI Press.

Fogg, P. C. M. (2019) 'Identification and Characterization of a Direct Activator of a Gene Transfer Agent', *Nature Communications*, 10: 595.

——, Westbye, A. B., and Beatty, J. T. (2012) 'One for All or All for One: Heterogeneous Expression and Host Cell Lysis Are Key to Gene Transfer Agent Activity in *Rhodobacter capsulatus*', *Plos One*, 7: e43772.

Fokine, A., and Rossmann, M. G. (2014) 'Molecular Architecture of Tailed Double-Stranded DNA Phages', *Bacteriophage*, 4: e28281.

Fu, Y. et al. (2010) 'High Diversity of *Rhodobacterales* in the Subarctic North Atlantic Ocean and Gene Transfer Agent Protein Expression in Isolated Strains', *Aquatic Microbial Ecology*, 59: 283–93.

Grull, M., Mulligan, M., and Lang, A. (2018) 'Small Extracellular Particles with Big Potential for Horizontal Gene Transfer: Membrane Vesicles and Gene Transfer Agents', *FEMS Microbiol Lett*, 365: fny192.

Henikoff, S., and Henikoff, J. G. (1993) 'Performance Evaluation of Amino Acid Substitution Matrices', *Proteins: Structure, Function, and Genetics*, 17: 49–61.

——, and —— (1994) 'Position-Based Sequence Weights', *Journal of Molecular Biology*, 243: 574–8.

Hoang, D. T. et al. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.

Hynes, A. P. et al. (2012) 'DNA Packaging Bias and Differential Expression of Gene Transfer Agent Genes within a Population during Production and Release of the *Rhodobacter capsulatus* Gene Transfer Agent, RcGTA', *Molecular Microbiology*, 85: 314–25.

—— et al. (2016) 'Functional and Evolutionary Characterization of a Gene Transfer Agent's Multilocus "Genome"', *Molecular Biology and Evolution*, 33: 2530–43.

Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.

Kogay, R. et al. (2019) 'Machine-Learning Classification Suggests That Many Alphaproteobacterial Prophages May instead Be Gene Transfer Agents', *Genome Biology and Evolution*, 11: 2941–53.

Kogay, R. et al. (2020) 'Selection for Reducing Energy Cost of Protein Production Drives the GC Content and Amino Acid Composition Bias in Gene Transfer Agents', *MBio*, 11: e01206–20.

Lang, A. S., and Beatty, J. T. (2000) 'Genetic Analysis of a Bacterial Genetic Exchange Element: The Gene Transfer Agent of *Rhodobacter capsulatus*', *Proceedings of the National Academy of Sciences of the United States of America* 97: 859–64.

——, Westbye, A. B., and Beatty, J. T. (2017) 'The Distribution, Evolution, and Roles of Gene Transfer Agents in Prokaryotic Genetic Exchange', *Annual Review of Virology*, 4: 87–104.

Leavitt, J. C. et al. (2013) 'Function and Horizontal Transfer of the Small Terminase Subunit of the Tailed Bacteriophage Sf6 DNA Packaging Nanomotor', *Virology*, 440: 117–33.

Lu, S. et al. (2020) 'CDD/SPARCLE: The Conserved Domain Database in 2020', *Nucleic Acids Research*, 48: D265–8.

McDaniel, L. D. et al. (2010) 'High Frequency of Horizontal Gene Transfer in the Oceans', *Science*, 330: 50.

Merrill, B. D. et al. (2016) 'Software-Based Analysis of Bacteriophage Genomes, Physical Ends, and Packaging Strategies', *BMC Genomics*, 17: 679.

Nagao, N. et al. (2015) 'The Gene Transfer Agent-like Particle of the Marine Phototrophic Bacterium *Rhodovulum sulfidophilum*', *Biochemistry and Biophysics Reports*, 4: 369–74.

Nguyen, L.-T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.

Philosof, A. et al. (2017) 'Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota', *Current Biology*, 27: 1362–8.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments', *Plos One*, 5: e9490.

Rao, V. B., and Feiss, M. (2008) 'The Bacteriophage DNA Packaging Motor', *Annual Review of Genetics*, 42: 647–81.

——, and —— (2015) 'Mechanisms of DNA Packaging by Large Double-Stranded DNA Viruses', *Annual Review of Virology*, 2: 351–78.

Schrödinger, LLC. (2020). The PyMOL molecular graphics system, version 2.4.

Shakya, M., Soucy, S. M., and Zhaxybayeva, O. (2017) 'Insights into Origin and Evolution of α-Proteobacterial Gene Transfer Agents', *Virus Evolution*, 3: vex036.

Sherlock, D., Leong, J. X., and Fogg, P. C. M. (2019) 'Identification of the First Gene Transfer Agent (GTA) Small Terminase in *Rhodobacter capsulatus* and Its Role in GTA Production and Packaging of DNA', *Journal of Virology*, 93: e01328–19.

Söding, J. (2005) 'Protein Homology Detection by HMM–HMM Comparison', *Bioinformatics (Oxford, England)*, 21: 951–60.

Sokal, R. R., and Michener, C. D. (1958) 'A Statistical Method for Evaluating Systematic Relationships', *Univ Kansas Sci Bull*, 38: 1409–38.

Solioz, M., Yen, H. C., and Marris, B. (1975) 'Release and Uptake of Gene Transfer Agent by *Rhodopseudomonas capsulata*', *Journal of Bacteriology*, 123: 651–7.

Steinegger, M., and Söding, J. (2017) 'MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets', *Nature Biotechnology*, 35: 1026–8.

Westbye, A. B. et al. (2017a) 'The *Rhodobacter capsulatus* Gene Transfer Agent is Induced by Nutrient Depletion and the RNAP Omega Subunit', *Microbiology*, 163: 1355–63.

——, Beatty, J. T., and Lang, A. S. (2017b) 'Guaranteeing a Captive Audience: Coordinated Regulation of Gene Transfer Agent (GTA) Production and Recipient Capability by Cellular Regulators', *Current Opinion in Microbiology*, 38: 122–9.

Whelan, S., and Goldman, N. (2001) 'A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach', *Molecular Biology and Evolution*, 18: 691–9.

Wolf, Y. I. et al. (2018) 'Origins and Evolution of the Global RNA Virome', *MBio*, 9: e02329–18.

Yang, Z. (1994) 'Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods', *Journal of Molecular Evolution*, 39: 306–14.

Yang, Y. et al. (2017) 'A Novel Roseosiphophage Isolated from the Oligotrophic South China Sea', *Viruses*, 9: 109.

Yutin, N. et al. (2018) 'Discovery of an Expansive Bacteriophage Family That Includes the Most Abundant Viruses from the Human Gut', *Nature Microbiology*, 3: 38–46.

Zhao, H. et al. (2013) 'Structures of the Phage Sf6 Large Terminase Provide New Insights into DNA Translocation and Cleavage', *Proceedings of the National Academy of Sciences of the United States of America*, 110: 8075–80.

Zimmermann, L. et al. (2018) 'A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at Its Core', *Journal of Molecular Biology*, 430: 2237–43.